Supplementary Information for

A Roadmap for Natural Product Discovery Based on Large-Scale Genomics and Metabolomics

James R. Doroghazi[†], Jessica C. Albright[†], Anthony W. Goering, Kou-San Ju, Robert R. Haines, Konstantin A. Tchalukov, David P. Labeda, Neil L. Kelleher[*], and William W. Metcalf[*]

† Authors contributed equally to this study

* Corresponding authors: William W. Metcalf, metcalf@illinois.edu; Neil L. Kelleher, n-kelleher@northwestern.edu

Supplementary Results

Method	PKS I	NRPS	PKS II	Lanthipeptides	NIS	TOMM
Extrapolation	8194	5757	1620	620	544	414
Chao 1	8299	5766	1654	621	478	413
ACE	8506	6190	1084	768	546	468

Supplementary Table 1 Estimated numbers of GCFs encoded by *Actinobacteria*

This table shows estimates from all three methods used on the same data for estimating the total number of gene clusters families. Extrapolation was performed out to 15000 genomes sampled, whereas Chao1 and ACE estimates are for the total number in each class based on the provided data. All methods were employed in the program estimateS [1]. These estimates only apply to the *Actinobacteria*.

Method	**PKS I**	**NRPS**	**PKS II**	**Lanthipeptides**	**NIS**	**TOMM**
Extrapolation	7527	4421	1373	492	403	355
Chao 1	7550	4431	1376	493	404	355
ACE	7715	5176	924	607	337	422

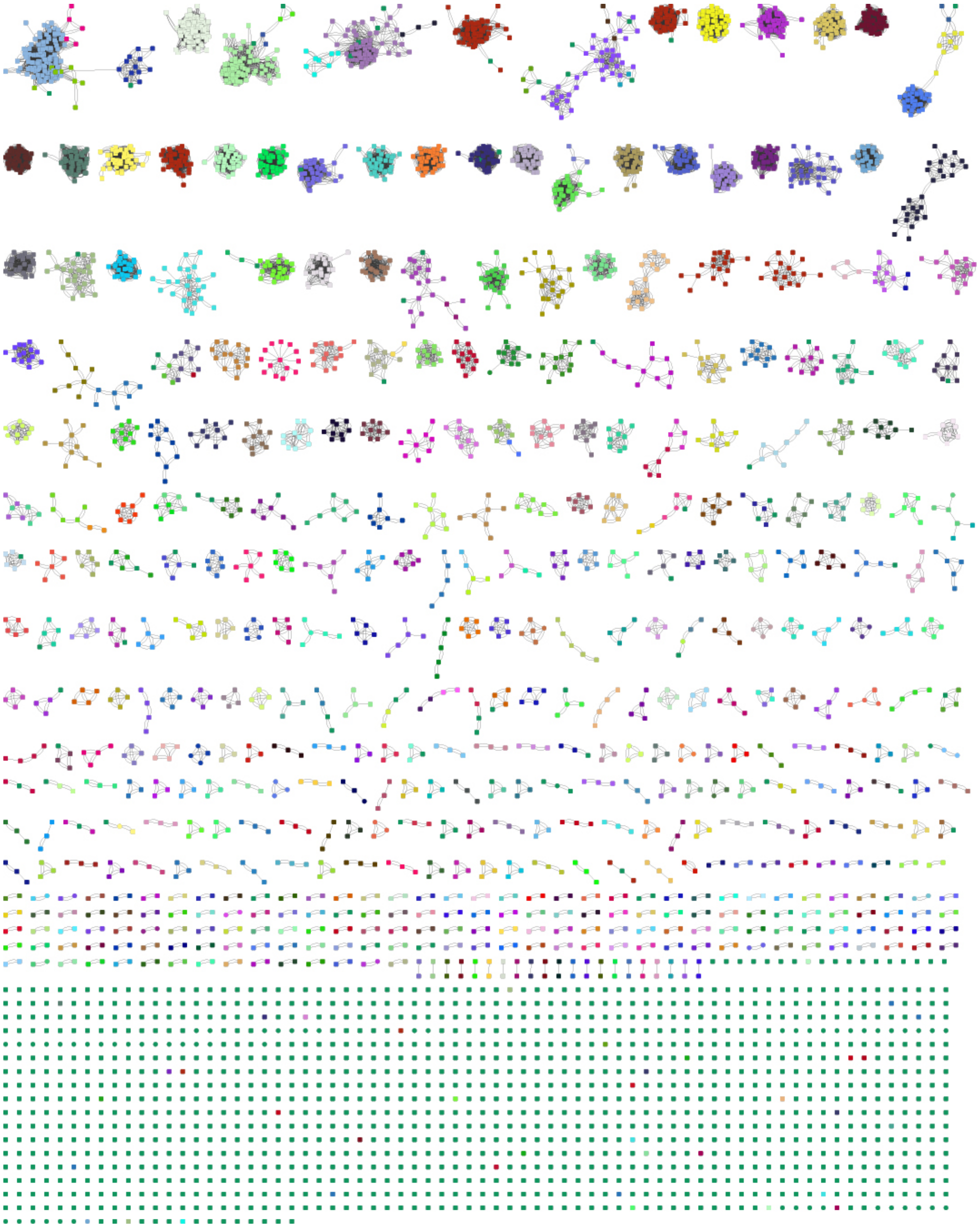
Supplementary Table 2 Estimated numbers of GCFs encoded by NP-rich *Actinobacteria*

This table shows estimates from all three methods used on the NP-rich genomic data for estimating the total number of gene clusters families. Only genomes found in Figure 1 between and including *Streptomycetales* to *Pseudonocardiales* are used to reduce any potential bias introduced by taxa not typically used for natural product discovery. Extrapolation was performed out to 15000 genomes sampled, whereas Chao1 and ACE estimates are for the total number in each class based on the provided data. All methods were employed in the program estimateS [1]. These estimates only apply to the subset of *Actinobacteria* described above.

Compound	UV max (literature)	UV max (observed)
Oxytetracycline	353	354
Proferrioxamine D2	435	433
Homononactyl nonactoate	Not reported	NA
Benarthin	252	254
Griseobactin	357	355
Kirromycin	233	--
Actinomycin D	442	441
Desertomycin A	223	225
Pyridomycin	227	228
Rimocidin amide	Not reported	NA
Enterocin	250	247
Chlortetracycline	224	226

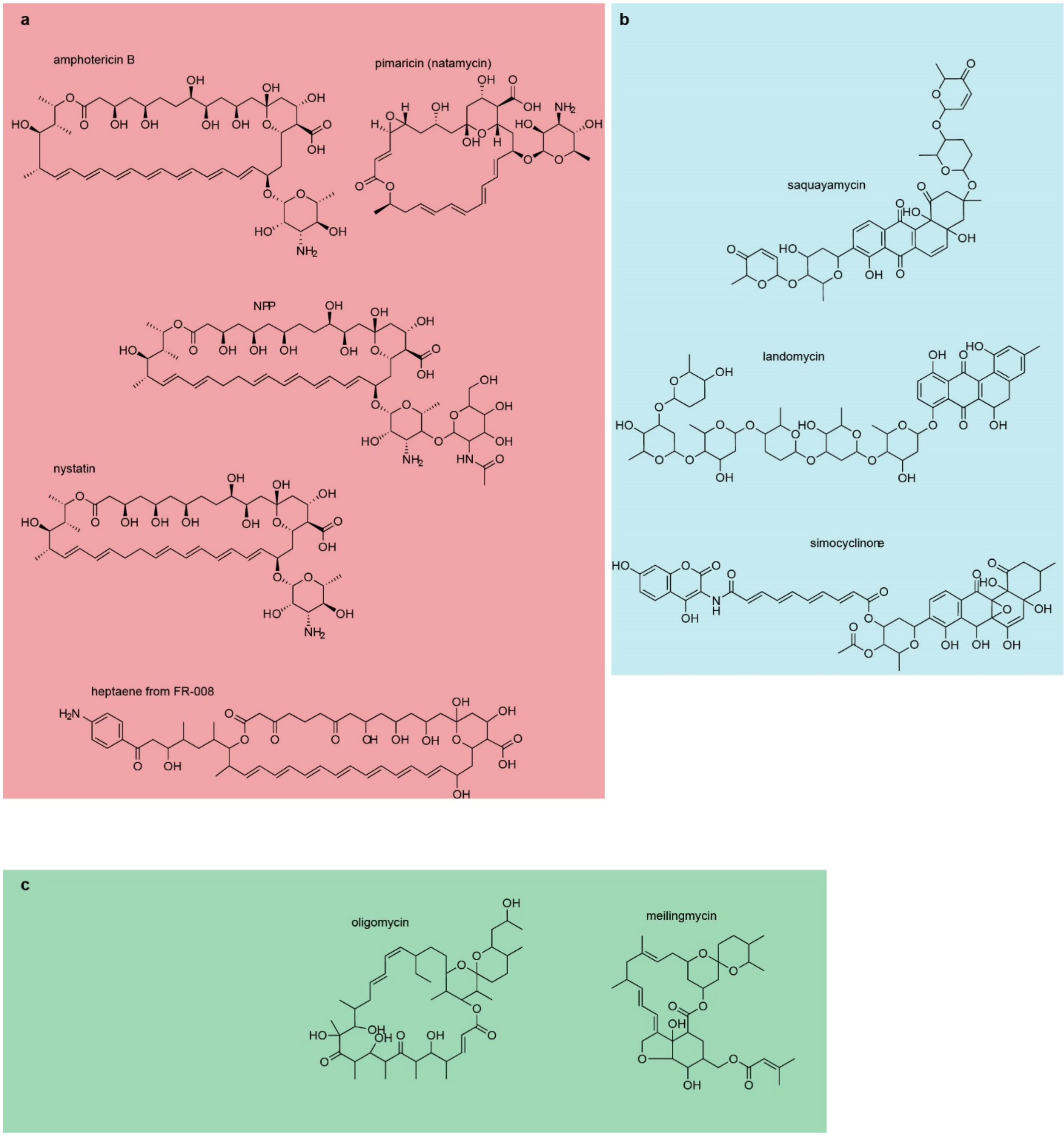
Supplementary Table 3 UV data to confirm mass-based compound identifications

Compounds identified via accurate intact mass were confirmed by comparing experimental UV absorption data to published literature values obtained from Dictionary of Natural Products. Kirromycin was present in only one strain, and co-eluted with several other high abundance species, so UV absorption data could not be obtained for this compound.



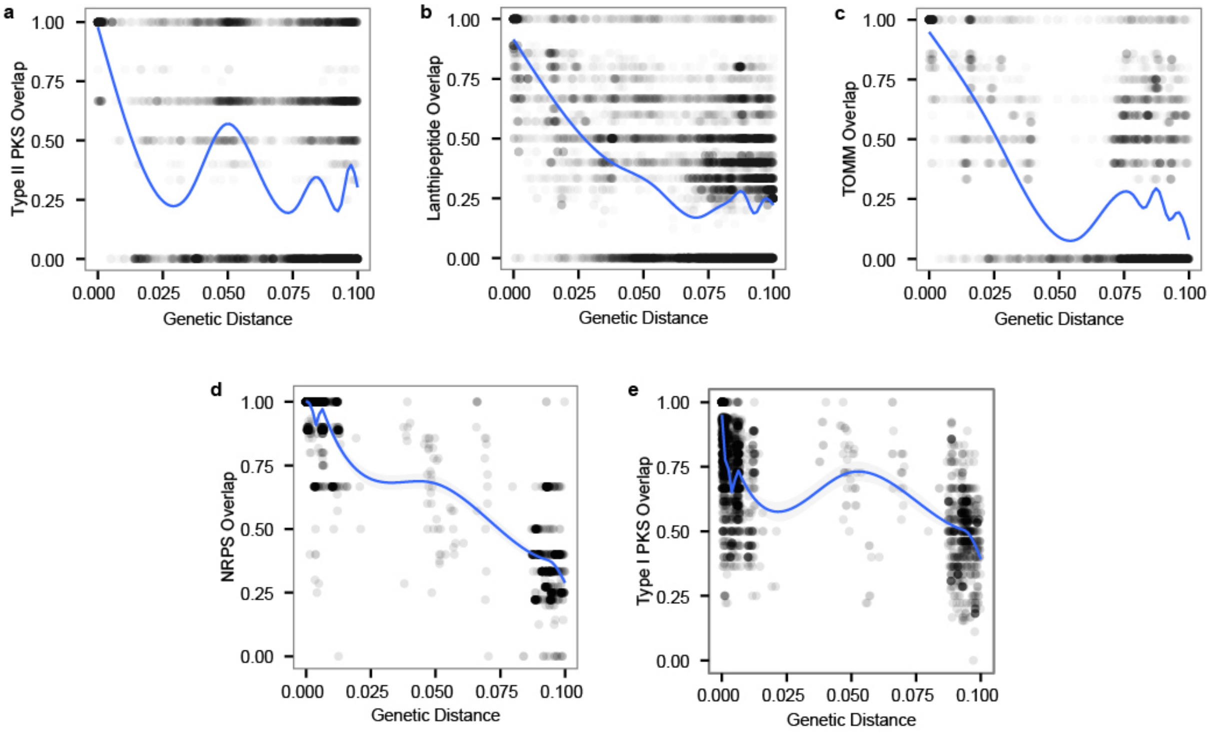
Supplementary Figure 1 The network of NRPS gene clusters

Each node in the network represents one gene cluster, edges are created based on automated scoring and manual curation, and a gene cluster family is defined as all gene clusters that can be connected by a path of any length. Node colors are based on density-based clustering. Gene cluster diagrams for each family can be accessed via the associated website (www.igb.illinois.edu/labs/metcalf/gcf) by clicking on the indicated node. To access the gene cluster diagrams, simply proceed to the biosynthetic class of interest through the Class dropdown menu. The Cytoscape networks described above are interactive, and clicking on any node opens a gene cluster diagram including all gene clusters in a GCF. All homologous genes present are highlighted on mouseover of any member of the homology group. The amino acid sequence can be accessed by clicking on any gene as well as a link to BLAST the same sequence. Boilerplate HTML5 was used for the creation of the site, including a template from: http://twitter.github.com/bootstrap/examples/hero.html.



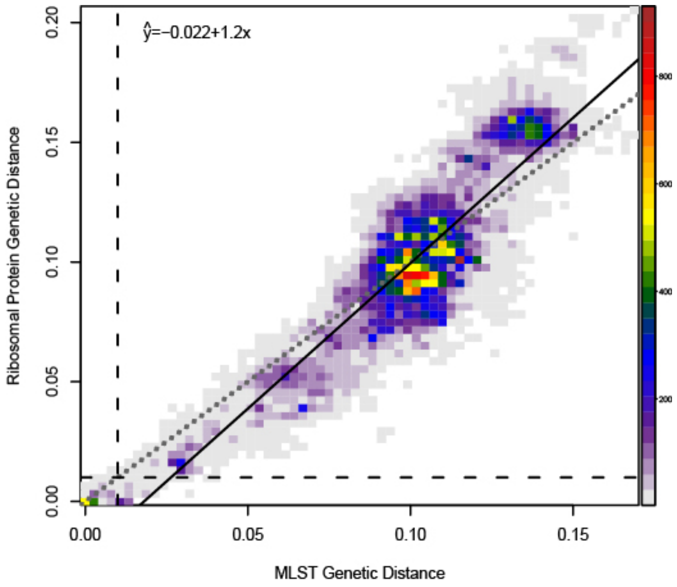
Supplementary Figure 2 Similar structures grouped into a GCF

Similar polyketides associated with NPGCs in (a) PKS_I_90, (b) PKS_II_49, and (c) PKS_I_1.



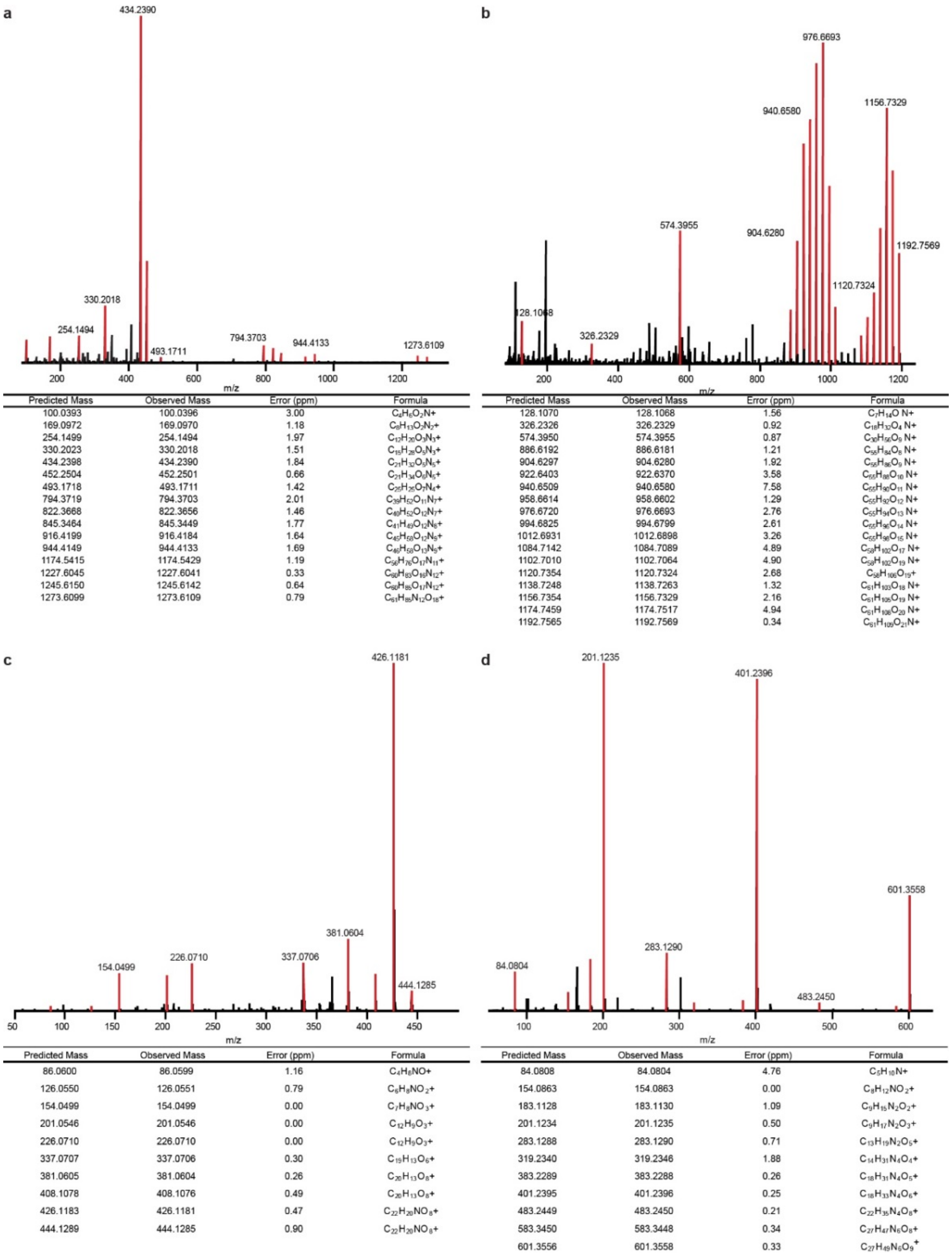
Supplementary Figure 3 GCF conservation across phylogenetic distance

GCF conservation is shown plotted against ribosomal protein distance for (a-c) every pair of natural product rich genomes between *Streptomycetales* and *Pseudonocardiales* in Figure 1a or (d-e) every pair of *Mycobacterium* genomes. Conservation of (a) type II PKS cluster, (b) lanthipeptide, and (c) TOMM, (d) NRPS, and (e) type I PKS GCFs plotted against ribosomal protein distance.



Supplementary Figure 4 Comparison of ribosomal protein and MLST evolutionary rates

Pairwise ribosomal protein amino acid distance is calculated based on identity and plotted against the nucleotide distance for the five gene multilocus sequence typing data set used for *Streptomyces* phylogenetic systematics. The legend on the right shows the color corresponding to the number of counts in each square. The dotted line goes through the intercept and has a slope of 1. The equation for the best fit line (solid) is shown. Multiple studies have shown that the species cutoff in *Streptomyces* is slightly less than 1%, from 0.7-0.8% [2-4]. The lines corresponding to 1% difference in sequence for each data set are shown as dashed lines.



Predicted Mass	Observed Mass	Error (ppm)	Formula
100.0393	100.0396	3.00	$C_4H_6O_2N$+
169.0972	169.0970	1.18	$C_8H_{13}O_2N_2$+
254.1499	254.1494	1.97	$C_{12}H_{20}O_3N_3$+
330.2023	330.2018	1.51	$C_{15}H_{28}O_5N_3$+
434.2398	434.2390	1.84	$C_{21}H_{32}O_5N_5$+
452.2504	452.2501	0.66	$C_{21}H_{34}O_6N_5$+
493.1718	493.1711	1.42	$C_{25}H_{25}O_7N_4$+
794.3719	794.3703	2.01	$C_{39}H_{52}O_{11}N_7$+
822.3668	822.3656	1.46	$C_{40}H_{52}O_{12}N_7$+
845.3464	845.3449	1.77	$C_{41}H_{49}O_{12}N_8$+
916.4199	916.4184	1.64	$C_{45}H_{58}O_{12}N_9$+
944.4149	944.4133	1.69	$C_{46}H_{58}O_{13}N_9$+
1174.5415	1174.5429	1.19	$C_{56}H_{76}O_{17}N_{11}$+
1227.6045	1227.6041	0.33	$C_{60}H_{83}O_{16}N_{12}$+
1245.6150	1245.6142	0.64	$C_{60}H_{85}O_{17}N_{12}$+
1273.6099	1273.6109	0.79	$C_{61}H_{85}N_{12}O_{18}$+

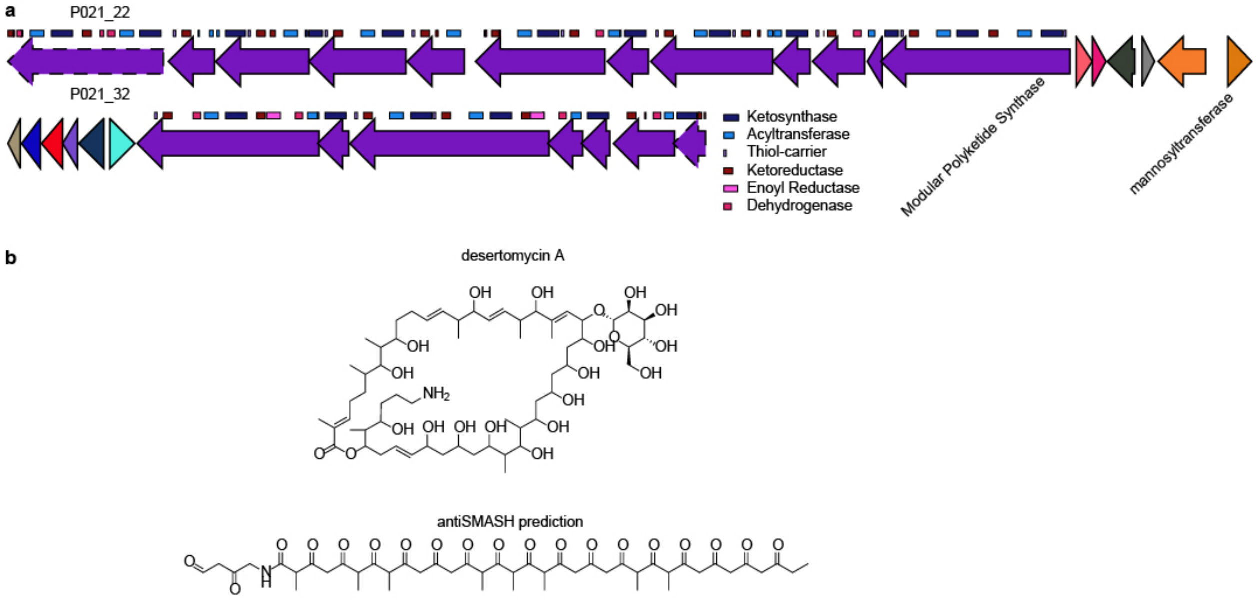
Predicted Mass	Observed Mass	Error (ppm)	Formula
128.1070	128.1068	1.56	$C_7H_{14}O$ N+
326.2326	326.2329	0.92	$C_{18}H_{32}O_4$ N+
574.3950	574.3955	0.87	$C_{30}H_{56}O_9$ N+
886.6192	886.6181	1.21	$C_{55}H_{84}O_8$ N+
904.6297	904.6280	1.92	$C_{55}H_{86}O_9$ N+
922.6403	922.6370	3.58	$C_{55}H_{88}O_{10}$ N+
940.6509	940.6580	7.58	$C_{55}H_{90}O_{11}$ N+
958.6614	958.6602	1.29	$C_{55}H_{92}O_{12}$ N+
976.6720	976.6693	2.76	$C_{55}H_{94}O_{13}$ N+
994.6825	994.6799	2.61	$C_{55}H_{96}O_{14}$ N+
1012.6931	1012.6898	3.26	$C_{55}H_{98}O_{15}$ N+
1084.7142	1084.7089	4.89	$C_{58}H_{102}O_{17}$ N+
1102.7010	1102.7064	4.90	$C_{58}H_{102}O_{19}$ N+
1120.7354	1120.7324	2.68	$C_{58}H_{106}O_{19}$+
1138.7248	1138.7263	1.32	$C_{61}H_{103}O_{18}$ N+
1156.7354	1156.7329	2.16	$C_{61}H_{105}O_{19}$ N+
1174.7459	1174.7517	4.94	$C_{61}H_{108}O_{20}$ N+
1192.7565	1192.7569	0.34	$C_{61}H_{109}O_{21}N$+

Predicted Mass	Observed Mass	Error (ppm)	Formula
86.0600	86.0599	1.16	C_4H_8NO+
126.0550	126.0551	0.79	$C_6H_8NO_2$+
154.0499	154.0499	0.00	$C_7H_8NO_3$+
201.0546	201.0546	0.00	$C_{12}H_9O_3$+
226.0710	226.0710	0.00	$C_{12}H_9O_3$+
337.0707	337.0706	0.30	$C_{19}H_{13}O_6$+
381.0605	381.0604	0.26	$C_{20}H_{13}O_8$+
408.1078	408.1076	0.49	$C_{20}H_{13}O_8$+
426.1183	426.1181	0.47	$C_{22}H_{20}NO_8$+
444.1289	444.1285	0.90	$C_{22}H_{20}NO_8$+

Predicted Mass	Observed Mass	Error (ppm)	Formula
84.0808	84.0804	4.76	$C_5H_{10}N$+
154.0863	154.0863	0.00	$C_8H_{12}NO_2$+
183.1128	183.1130	1.09	$C_9H_{15}N_2O_2$+
201.1234	201.1235	0.50	$C_9H_{17}N_2O_3$+
283.1288	283.1290	0.71	$C_{13}H_{19}N_2O_5$+
319.2340	319.2346	1.88	$C_{14}H_{31}N_4O_4$+
383.2289	383.2288	0.26	$C_{18}H_{31}N_4O_5$+
401.2395	401.2396	0.25	$C_{18}H_{33}N_4O_6$+
483.2449	483.2450	0.21	$C_{22}H_{35}N_4O_8$+
583.3450	583.3448	0.34	$C_{27}H_{47}N_6O_8$+
601.3556	601.3558	0.33	$C_{27}H_{49}N_6O_9^+$

Supplementary Figure 5 Representative spectra from tandem MS compound identification

High resolution MS/MS spectrum of compound identified as (a) actinomycin G6, (b) desertomycin A, (c) oxytetracycline, and (d) nocardamine. Fragment ions in red are consistent with those predicted for each compound. Masses of predicted and observed fragment ions are listed below.



Supplementary Figure 6 The desertomycin biosynthetic gene cluster

(a) The predicted desertomycin biosynthetic gene cluster found through correlation of GCF and MS data. P021 is the locus tag for *Streptomyces alboflavus* NRRL B-2373. The domains found by pHMM are shown above the modular polyketide synthase genes. Two separate contigs from *Streptomyces alboflavus* NRRL B-2373 are included in gene cluster family PKS_I_18. It is possible that there is a missing piece of this cluster that was not assembled due to the draft status of the genome assembly. (b) The structure of desertomycin A is shown along with the predicted product of the gene cluster provided by antiSMASH [5]. The prediction was performed based on the concatenated nucleotide sequences of both clusters in the order of P021_32-P021_22.

Supplementary Data Set 1 A spreadsheet that lists the locus tags and corresponding organism for each genome in this study.

Supplementary Data Set 2 All gene cluster families that have at least one characterized natural product are listed in this Excel spreadsheet. The first column is the gene cluster family designation. The second column is the assigned cluster ID. Both of these are the same as found in Supplementary Data Set 1. The third column is the strain in which the gene cluster was found. For characterized gene clusters from Genbank, the title of the Genbank entry is given.

Supplementary Data Set 3 A spreadsheet containing raw MS data for the known compounds identified in this study. Each genome (column) is identified by the corresponding locus tag prefix, and each compound (row) is identified by common name, retention time and mass-to-charge ration (*m/z*). Each data cell contains the detected intensity of the given compound in the corresponding strain.

Supplementary Information References

1 Colwell, R. K. *et al.* Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *J. Plant Ecol.* **5**, 3-21 (2012).

2 Rong, X. Y., Guo, Y. P. & Huang, Y. Proposal to reclassify the *Streptomyces albidoflavus* clade on the basis of multilocus sequence analysis and DNA-DNA hybridization, and taxonomic elucidation of *Streptomyces griseus* subsp *solvifaciens*. *Syst. Appl. Microbiol.* **32**, 314-322 (2009).

3 Rong, X. Y. & Huang, Y. Taxonomic evaluation of the *Streptomyces griseus* clade using multilocus sequence analysis and DNA-DNA hybridization, with proposal to combine 29 species and three subspecies as 11 genomic species. *Int. J. Syst. Evol. Microbiol.* **60**, 696-703 (2010).

4 Rong, X. Y. & Huang, Y. Taxonomic evaluation of the *Streptomyces hygroscopicus* clade using multilocus sequence analysis and DNA-DNA hybridization, validating the MLSA scheme for systematics of the whole genus. *Syst. Appl. Microbiol.* **35**, 7-18 (2012).

5 Medema, M. H. *et al.* antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Res.* **39**, W339-W346 (2011).

Supplementary Note

The following characterized biosynthetic gene clusters were grouped together into gene cluster families (references are given where similarities of the compounds are discussed, all others are shown in Supplementary Figure 2): A-500359s, A-503083 [1]; C-1027, neocarzinostatin, maduropeptin [2]; zorbamycin, bleomycin, tallysomycin [3]; triostin, thiocoraline, SW-163 [4]; myxothiazol, melithiazol (in PKS I and NRPS networks) [5]; FK506 (2 clusters), FK520 (in PKS I and NRPS networks) [6]; balhimycin, teicoplanin, A40926, A47934 [7]; oligomycin, meilingmycin (Supplementary Fig. 2); herbimycin, geldanamycin (3 clusters) [8]; jerangolid, ambruticin [9]; BE-14106, ML-449 [10]; compactin, monacolin K [11]; nigericin, nanchangmycin [12]; megalomycin, erythromycin [13]; dihydrochalcomycin, anglomycin; AHBA (2 clusters), rifamycin [14]; NPP, amphotericin, pimaricin (3 clusters), heptaene macrolide from *Streptomyces* sp. FR-008, nystatin (Supplementary Fig. 2); lysolipin, rubrinomycin, griseorhodin [15]; saquayamycin Z, simocyclinone (2 clusters), landomycin (Supplementary Fig. 2); ravidomycin, chrysomycin [16]; anabaenopeptin (2 clusters); phosphinothricin tripeptide (2 clusters); nostopeptolide (2 clusters); microcystin (3 clusters, in PKS I and NRPS networks); AM-toxin (3 clusters); glycopeptidolipid (2 clusters, in PKS I and NRPS networks); epothilone (4 clusters, in PKS I and NRPS networks); tautomycetin (2 clusters); cylindrospermopsin (2 clusters); pyoluteorin (2 clusters); pactamycin (2 clusters); fumonisin (2 clusters); lasalocid (2 clusters); lactonamycin (2 clusters); aclacinomycin (2 clusters); SapB (2 clusters); microcyclamide (2 clusters); pyrrolomycin (2 clusters); siomycin, thiostrepton [17]; nocathiacin, nosiheptide [18]; pacidamycin, uridyl peptide antibiotic from *Streptomyces coeruleorubidus* strain AB1183F-64 [19].

Supplementary Note References

1 Muramatsu, Y. *et al.* A-503083 A, B, E and F, novel inhibitors of bacterial translocase I, produced by Streptomyces sp. SANK 62799. *J. Antibiot.* **57**, 639-646 (2004).

2 Gao, B. & Gupta, R. S. Phylogenetic framework and molecular signatures for the main clades of the phylum *Actinobacteria*. *Microbiol. Mol. Biol. Rev.* **76**, 66-112 (2012).

3 Whitman, W. B. *et al. Bergey's manual® of systematic bacteriology*. Vol. 5 (Springer, 2012).

4 Jones, A. L. & Goodfellow, M. in *Bergey's manual® of systematic bacteriology* Vol. 5 (eds William B Whitman *et al.*) 437-464 (Springer, 2012).

5 Weinig, S., Hecht, H.-J., Mahmud, T. & Müller, R. Melithiazol biosynthesis: further insights into myxobacterial PKS/NRPS systems and evidence for a new subclass of methyl transferases. *Chem. Biol.* **10**, 939-952 (2003).

6 Ludwig, W. *et al.* in *Bergey's manual® of systematic bacteriology* Vol. 5 (eds William B Whitman *et al.*) 1-28 (Springer, 2012).

7 Global tuberculosis report 2012. (World Health Organization, Geneva, Switzerland, 2012).

8 Laber, B., Lindell, S. D. & Pohlenz, H.-D. Inactivation of Escherichia coli threonine synthase by dl-Z-2-amino-5-phosphono-3-pentenoic acid. *Arch. Microbiol.* **161**, 400-403 (1994).

9 Julien, B., Tian, Z.-Q., Reid, R. & Reeves, C. D. Analysis of the Ambruticin and Jerangolid Gene Clusters of *Sorangium cellulosum* Reveals Unusual Mechanisms of Polyketide Biosynthesis. *Chem. Biol.* **13**, 1277-1286 (2006).

10 Jørgensen, H. *et al.* Insights into the evolution of macrolactam biosynthesis through cloning and comparative analysis of the biosynthetic gene cluster for a novel macrocyclic lactam, ML-449. *Appl. Environ. Microbiol.* **76**, 283-293 (2010).

11 Chen, Y.-P. *et al.* Cloning and characterization of monacolin K biosynthetic gene cluster from Monascus pilosus. *J. Agric. Food Chem.* **56**, 5639-5646 (2008).

12 Harvey, B. M. *et al.* Insights into Polyether Biosynthesis from Analysis of the Nigericin Biosynthetic Gene Cluster in *Streptomyces* sp. DSM4137. *Chem. Biol.* **14**, 703-714 (2007).

13 Sarker, S. D., Latif, Z. & Gray, A. I. *Natural products isolation*. Vol. 20 (Springer, 2005).

14 Evans, B. S. *et al.* Discovery of the Antibiotic Phosacetamycin via a New Mass Spectrometry-Based Method for Phosphonic Acid Detection. *ACS Chem. Biol.* **8**, 908-913 (2013).

15 Kudo, F., Yonezawa, T., Komatsubara, A., Mizoue, K. & Eguchi, T. Cloning of the biosynthetic gene cluster for naphthoxanthene antibiotic FD-594 from Streptomyces sp. TA-0256. *J. Antibiot.* **64**, 123-132 (2010).

16 Kharel, M. K., Nybo, S. E., Shepherd, M. D. & Rohr, J. Cloning and characterization of the ravidomycin and chrysomycin biosynthetic gene clusters. *ChemBioChem* **11**, 523-532 (2010).

17 Nishimura, H. *et al.* Siomycin, a new thiostrepton-like antibiotic. *J. Antibiot.* **14**, 255-263 (1961).

18 Wei, M., Deng, J., Wang, S., Liu, N. & Chen, Y. A simple reverse genetics approach to elucidating the biosynthetic pathway of nocathiacin. *Biotechnol. Lett.* **33**, 585-591 (2011).

19 Zhang, W., Ostash, B. & Walsh, C. Identification of the biosynthetic gene cluster for the pacidamycin group of peptidyl nucleoside antibiotics. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 16828-16861 (2010).